

## 周报（2014.11.03-2014.11.09）

1. 本周周一讨论了目前有的温州数据的可用性。针对大家的讨论后面几天工作主要围绕寻找可用的数据展开。杨哲和李嘉华都对数据进行了探索。

对于浙江省所有人的上网数据，由于这些数据只有每个人访问了什么网站，也就是我们只能获得 http 请求中 get 请求的情况，所以我们无法直接得到每个用户对应的微博账号。不过对于 url 以 weibo.cn/?gsid= 开始的是用户的登陆请求，像 weibo.cn/1914953542/ 这样的后面有一串数字 id 的是用户微博网页浏览。我们可以认为微博的一般访问模式是先登陆账号，然后是自己的微博主页，后面可能访问其他人的微博主页。不过综合来看访问自己的应该会稍微多一些。我和杨哲通过一个用户的浏览时的 repost 情况，并查看相关微博通过猜测并验证了该用户的微博账号。目前如果我们想从这个数据中获得每个手机用户的微博账号只能通过这种猜测并验证的方法进行。这部分的程序杨哲还在写，下周的前两天应该差不多写完，不过由于该数据的数据量比较大，600 多 G，程序运行的时间可能比较久。这些等程序写出来会再认真考虑。

关于其他的当时讨论的数据，我们后面从网上没有找到可用的。

关于该项目使用哪些数据，后面又多次看了 palantir 的相关视频，发现它的 demo 里面一般涉及到的异构数据集也不是很多，有的也只有 2 个，大部分也是只有三四个。所以我在想是否目前我们有的温州这些数据集就可以使用。当然如果我们决定了用这个数据集，后面也会一直考虑是否有新的数据集可以加进来。

2. 关于本体构建这部分，我和王琦一直还在看相关的材料。找到了一些零零散散的小的相关的本体。关于这部分，我感觉根据我们目前的需求，我们既不需要构建一个很全的顶级本体也不需要构建一个很权威的领域本体，只需要根据实际的需求构建一个相对比较全的、我们程序员使用起来比较方便的、逻辑不太抽象的、用户也容易理解的本体。如果是这样的话，我们可以参考一些现有的材料，从现有的本体构建。最近感觉这部分的材料已经找的还可以，后面的构建应该不会太困难。<http://notes.3kbo.com/> 这个网站我认为关于本体总结的不错，也收集了很多材料，从具体的领域、应用、已有的本体等方面进行了总结。本体的构建下星期会继续。

下周工作：

1. 数据部分：杨哲继续进行上网数据的处理，尽快找到手机用户和微博账号的对应关系。
2. 本体的构建这部分也会继续进行。
3. 尽快跟大家讨论确定数据。